Facial
~~Biometrics Human~~ Examiner Overview

Dr. P. Jonathon Phillips

National Institute of Standards and Technology

PNAS Proceedings of the National Academy of Sciences of the United States of America

Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms

P. Jonathon Phillips, Amy N. Yates
National Institute of Standards and Technology

Ying Hu, Carina A. Hahn, Eilidh Noyes, Kelsey Jackson, Jacqueline Castro, Geraldine Jeckeln
The U of Texas at Dallas

Rajeev Ranjan, Swami Sankaranarayanan , Jun-Cheng Chen, Carlos D. Castillo, Rama Chellappa
U. of Maryland

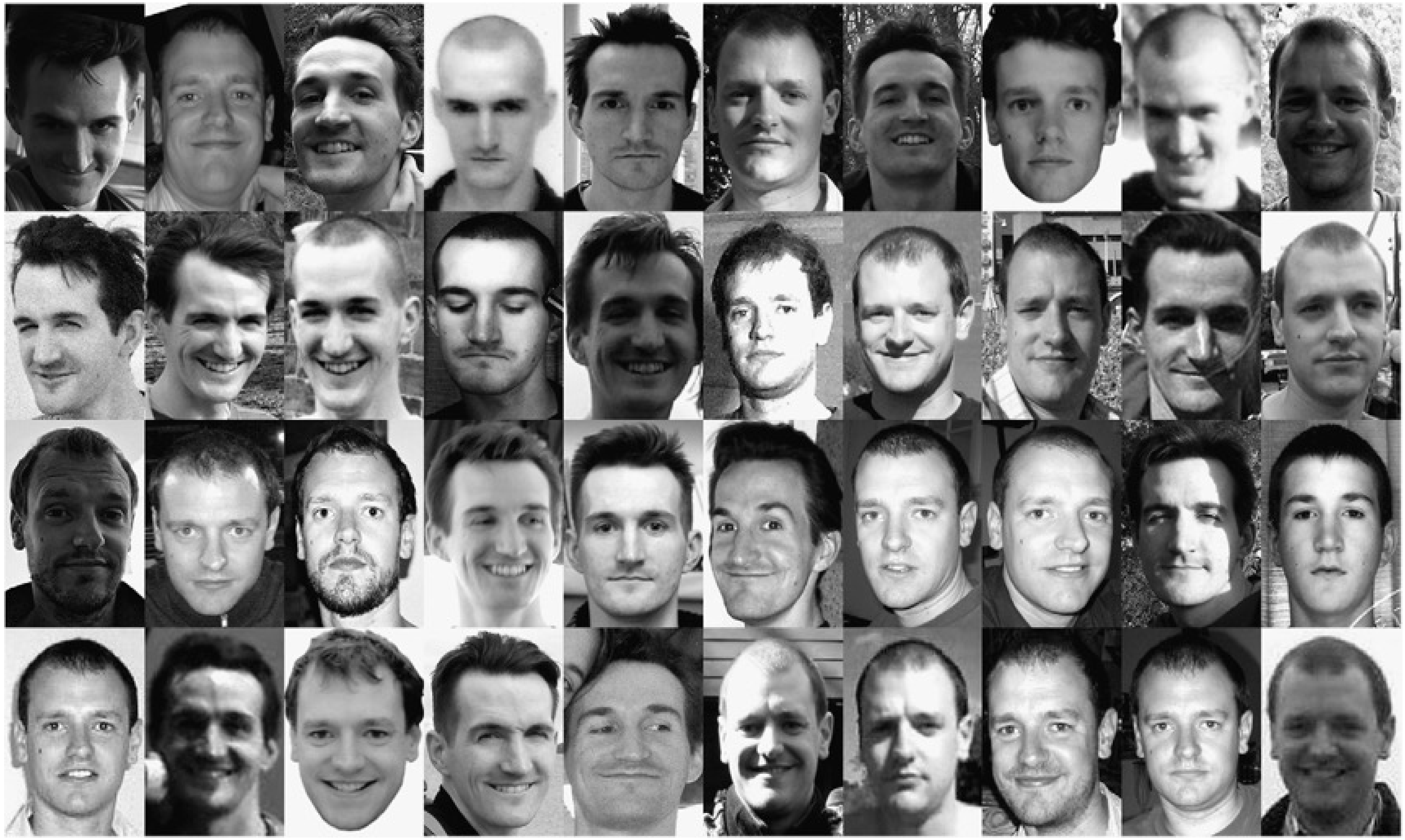
David White
UNSW - Sydney

Alice J. O'Toole
The U of Texas at Dallas

Pop Quiz

Who is this person?

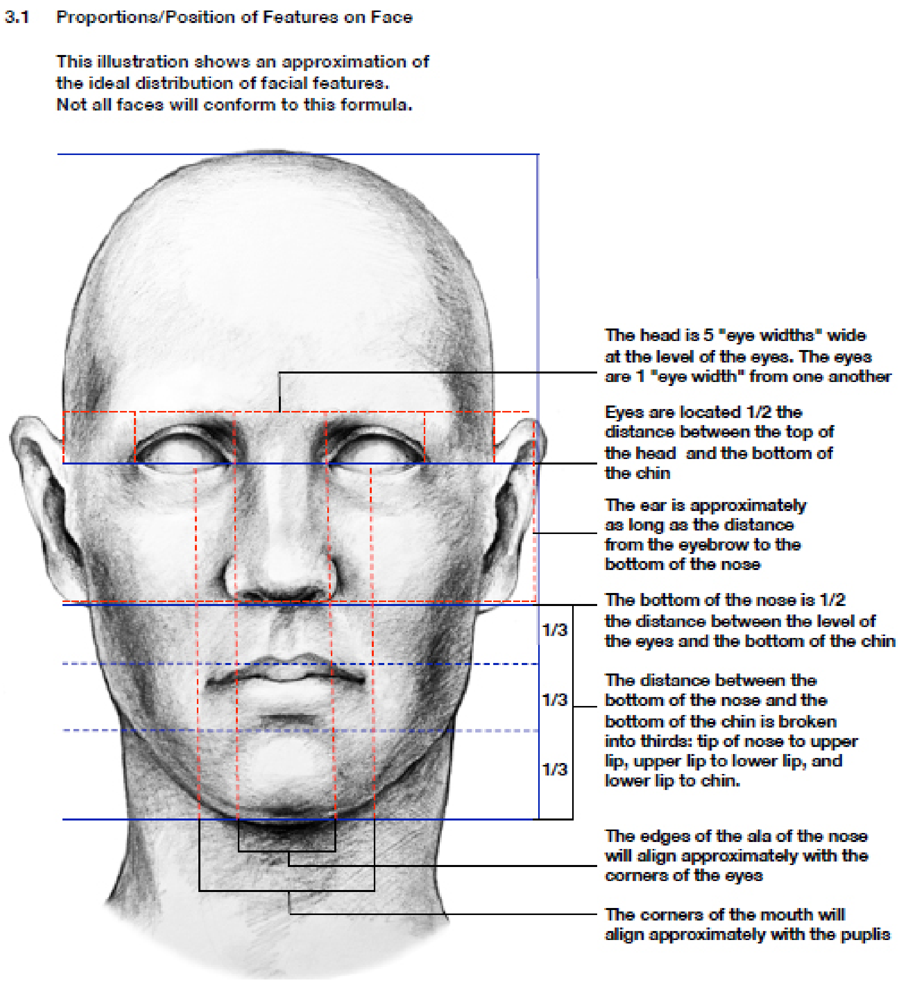
How many people?



Jenkins, White, Burton (2011)

What is a facial examiner?

- Comprehensive comparison of faces in images
- Write detailed reports
- Prepared to testify in court
- Extensive training (2-4 years)



What is a super-recognizer?

No consistent definition

Three criteria in the literature and media

Self reports

Employed as a super-recognizer

- London Met's Super-recognizer unit

Superior performance on a psychological face test

- Cambridge Face Memory Test
- Glasgow Face Matching Test

Noyes E, Phillips PJ, O'Toole AJ (2017) *What is a super-recogniser? Face Processing: Systems, Disorders, and Cultural Differences, eds Bindermann M, Megreya AM (Nova, New York), pp 173–201.*

Black box study

- Measure performance of Forensic Facial Examiners *in situ*.

- Examiners were allowed access to lab procedures, tools, methods, resources, and time schedule (more or less).

General rules

20 pairs of face images

- Pre-screened by humans and machines to be ***extremely*** challenging

7 point comparison scale

3 months to complete comparisons

Same-identity pair

Different-identity pair

Comparisons

+3	The observations strongly support that it is the same person
+2	The observations support that it is the same person
+1	The observations support to some extent that it is the same person
0	The observations support neither that it is the same person nor that it is different persons
-1	The observations support to some extent that it is not the same person
-2	The observations support that it is not the same person
-3	The observations strongly support that it is not the same person

Four subject groups ++

Facial forensic examiners (n=87, 5 continents)

- Examiners (n=57)
- Reviewers (n=30)

Super-recognizers (n=13)

Fingerprint examiners with no face experience (n=53)

Undergraduate Students (n=30)

Algorithms

Algorithms

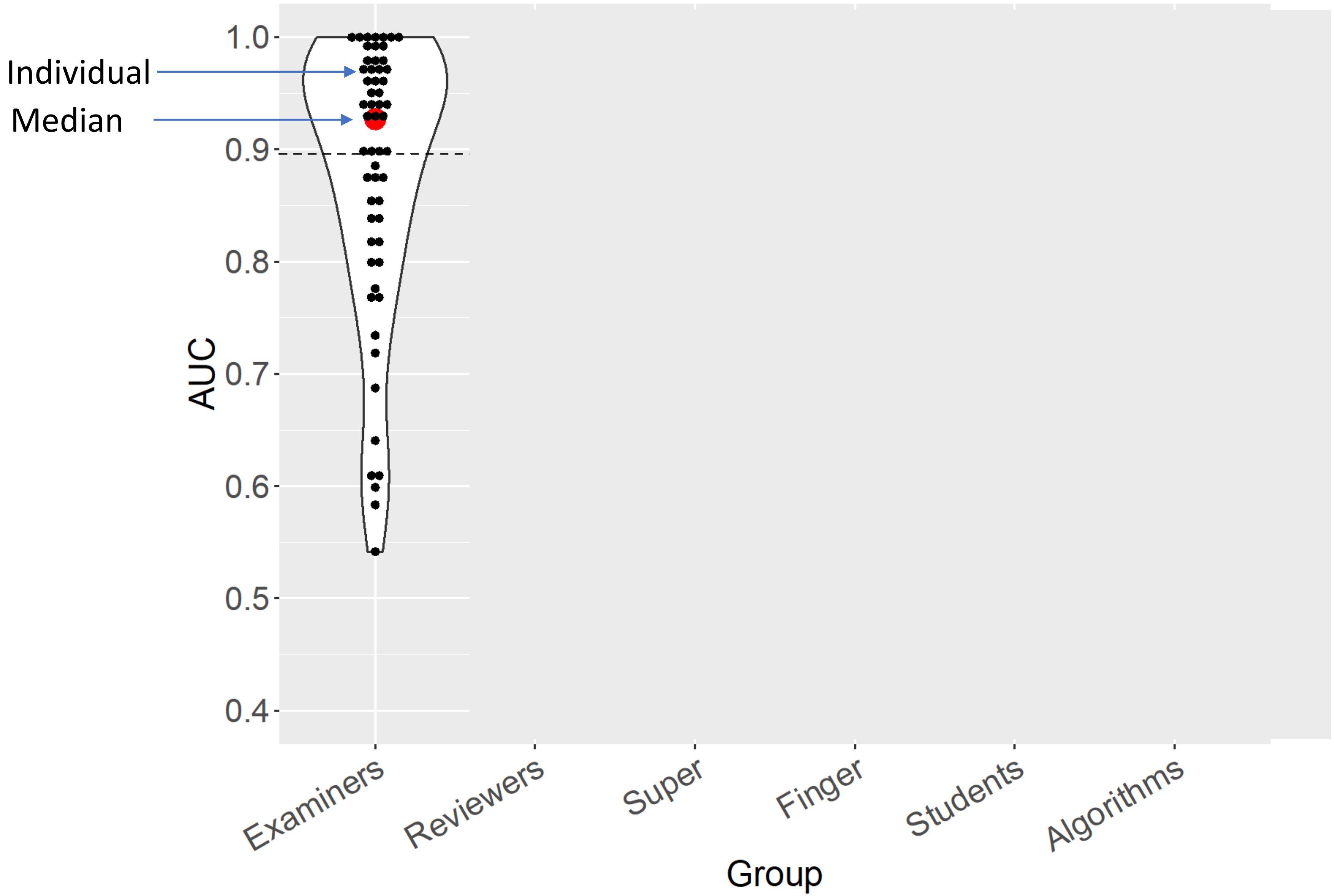
VGG-Face (A2015)

- Benchmark algorithm from Oxford
- Deep convolutional neural network (DCNN) based

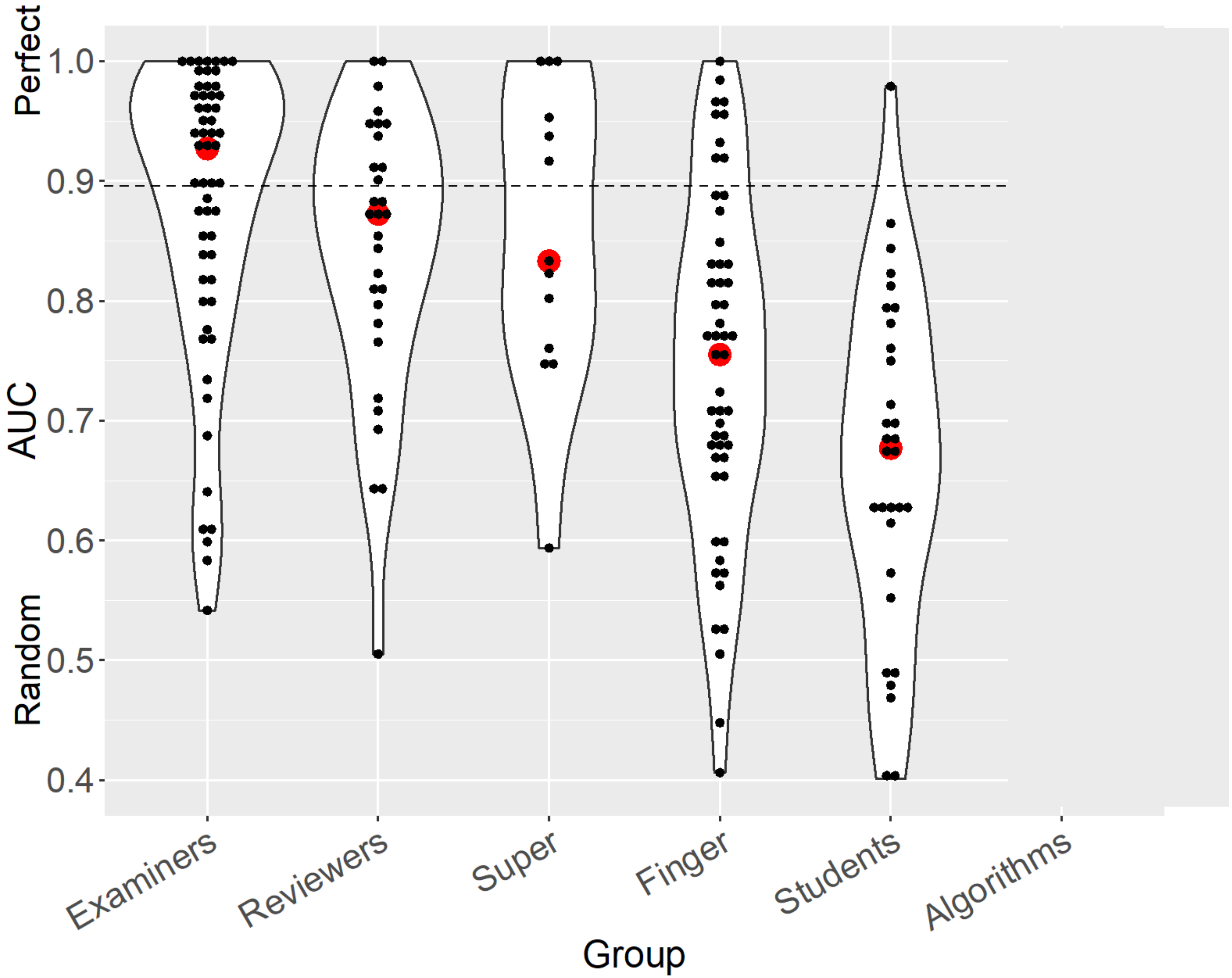
U of Maryland

- Rama Chellappa's group
- 3 algorithms (A2016, A2017a, A2017b)
- Deep convolutional neural network (DCNN) based

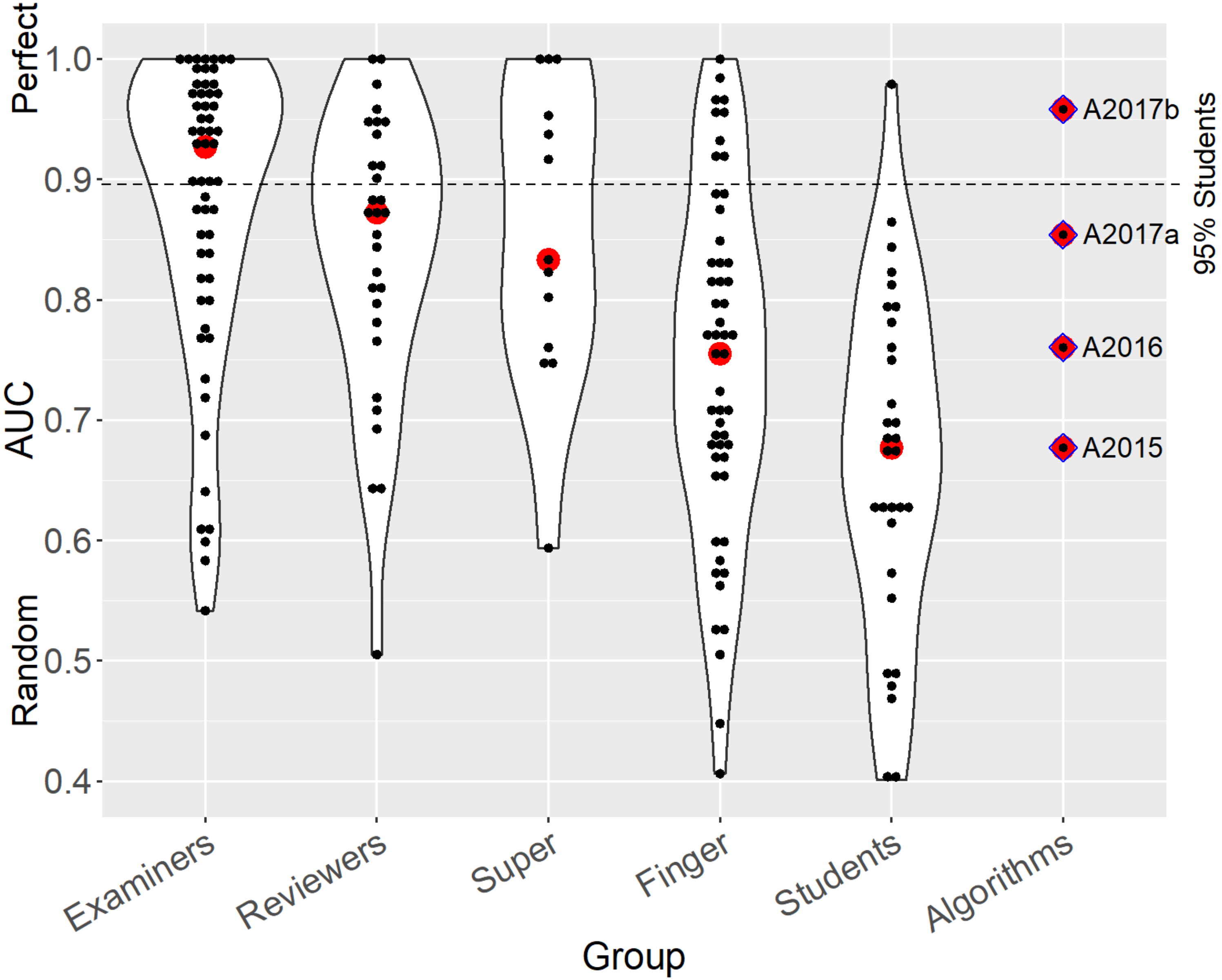
Comparison across groups



Comparison Across Groups



Comparison across groups

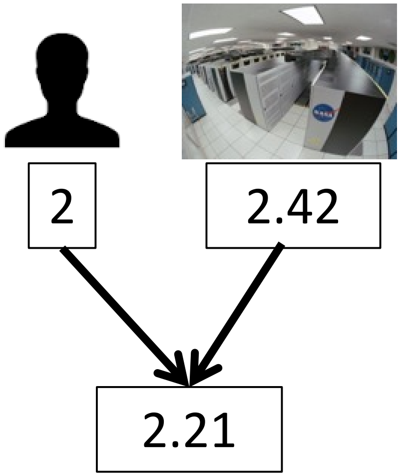


Fusing examiners and algorithms

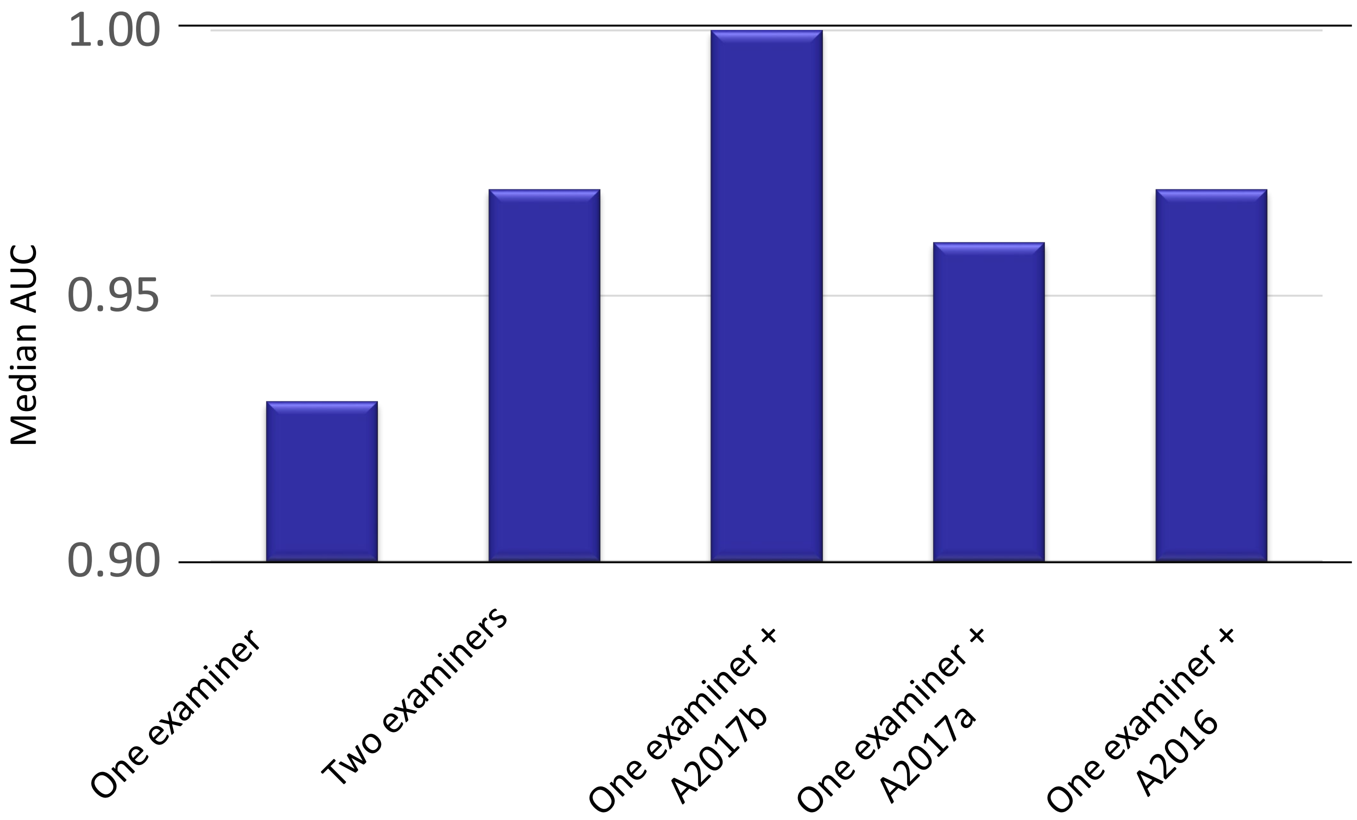
NIST

Rescale algorithm scores

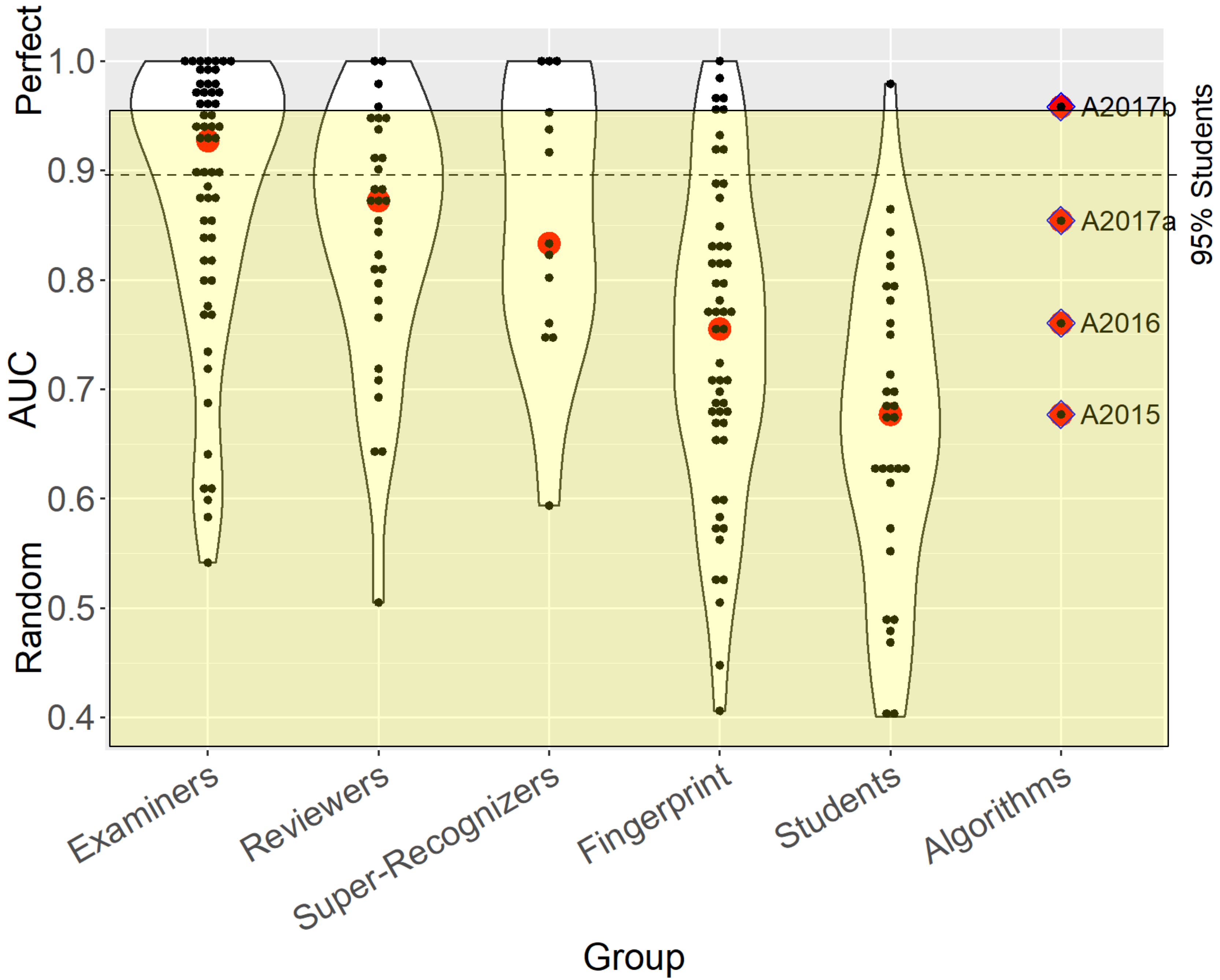
Fusion by averaging



Fusing examiners and algorithms



Implications of fusing humans and algorithms



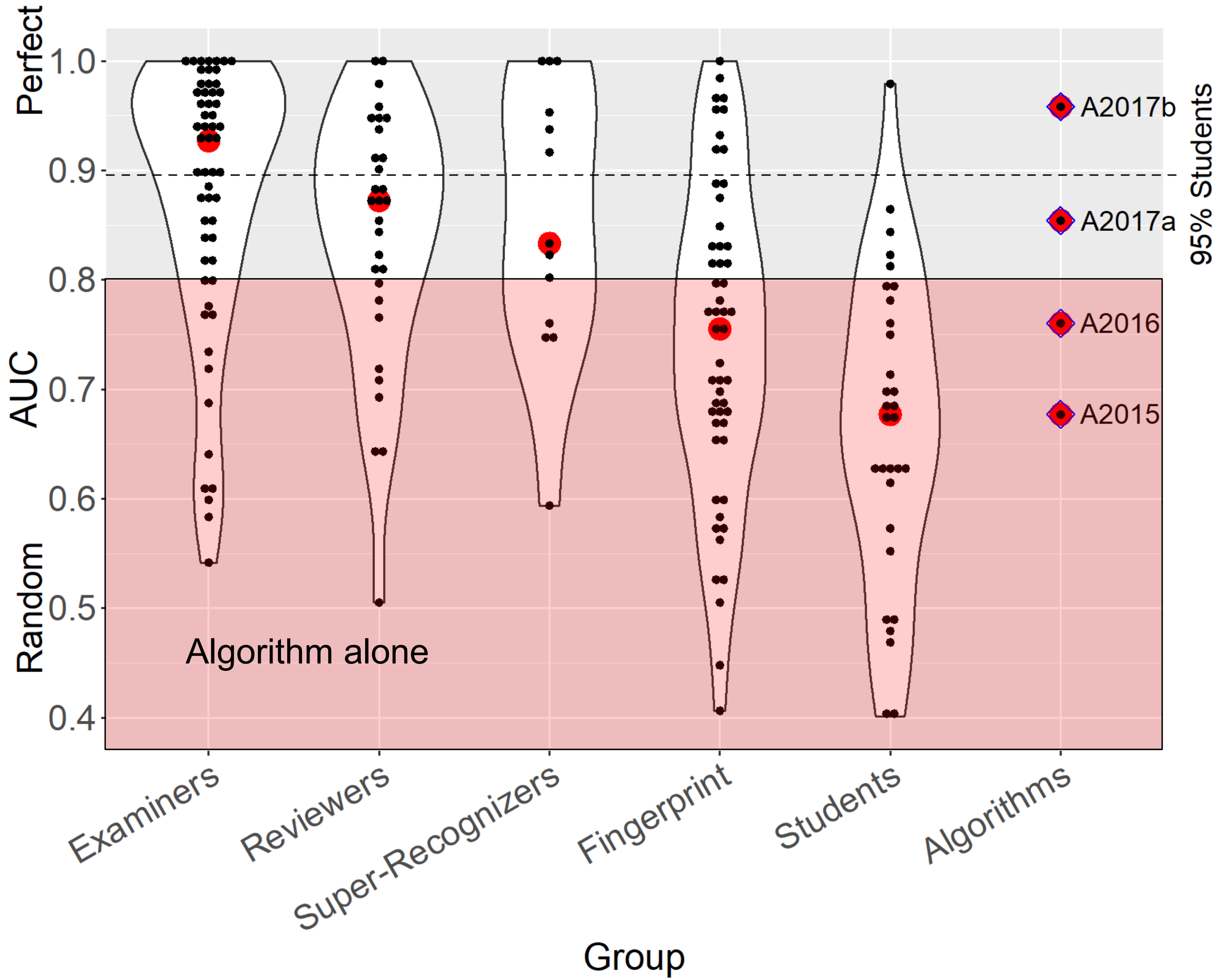
Perfect
1.0
0.9
0.8
AUC
0.7
0.6
Random
0.5
0.4
A2017b
A2017a
A2016
A2015
95% Students
Examiners
Reviewers
Super-Recognizers
Fingerprint
Students
Algorithms
Group

To fuse or not to fuse?

Fuse judgments between people and algorithm (A2017b)

Determine which is most accurate in following scenarios:

- ~~Human alone~~
- Algorithm alone (A2017b)
- Fusing human & algorithm



Perfect
Random
AUC
1.0
0.9
0.8
0.7
0.6
0.5
0.4
95% Students
A2017b
A2017a
A2016
A2015
Algorithm alone
Examiners
Reviewers
Super-Recognizers
Fingerprint
Students
Algorithms
Group

PNAS conclusions

- Facial examiners are significantly better than the general population

- No statistical difference among examiners, reviewers, and super-recognizers.
- Best algorithm is competitive with best humans
- Fusing human judgements is effective
- Performance optimized by fusing one facial examiner and A2017b.

Deeper understanding of examiners

- General face intelligence
  - Performance across race
  - Face memory
  - Disguises
- Item response theory for designing tests
- Comparison of examiners and super-recognizers

Black box road map

1. Perceptual test with students
2. Perceptual test with face professionals (e.g., face examiners)[1]
3. Black box test

[1]D. White, P.J. Phillips, C.A. Hahn, M. Hill, and A.J. O'Toole, "Perceptual expertise in forensic facial image comparison," *Proceedings of the Royal Society B*, 282(1814), 2015.

DOI: [10.1098/rspb.2015.1292](https://doi.org/10.1098/rspb.2015.1292)

Performance across races

Team

- University of Texas at Dallas
  - Prof. Alice J. O’Toole
  - Jacqueline Cavazos
  - Ying Hu
  - Geraldine Jeckeln
- University of Huddersfield (UK)
  - Dr. Eilidh Noyes
- NIST
  - Dr. Amy N. Yates
  - Dr. P. Jonathon Phillips
  - Dr. Carina A. Hahn

Why across races?

The other race effect (ORE) is when it is easier to recognize faces of one's own race than it is to recognize faces of a different race.

It is important to know how well face examiners perform across different races.

Test based on imagery in Phillips *et al.* (2011)[2], contains equal sets of Caucasian and East Asian faces.

[2]P. J. Phillips, F. Jiang, A. Narvekar, J. Ayyad, and A. J. O'Toole, "An other-race effect for face recognition algorithms," *ACM Trans. Appl. Percept.*, vol. 8, Feb. 2011. DOI: 10.1145/1870076.1870082

Example image pairs

+2 : Sure they are the same
+1 : Think they are the same
0 : Do not know
-1 : Think they are not the same
-2 : Sure they are not the same

Same Identity

Example Image Pairs

+2 : Sure they are the same
+1 : Think they are the same
0 : Do not know
-1 : Think they are not the same
-2 : Sure they are not the same

Different Identities

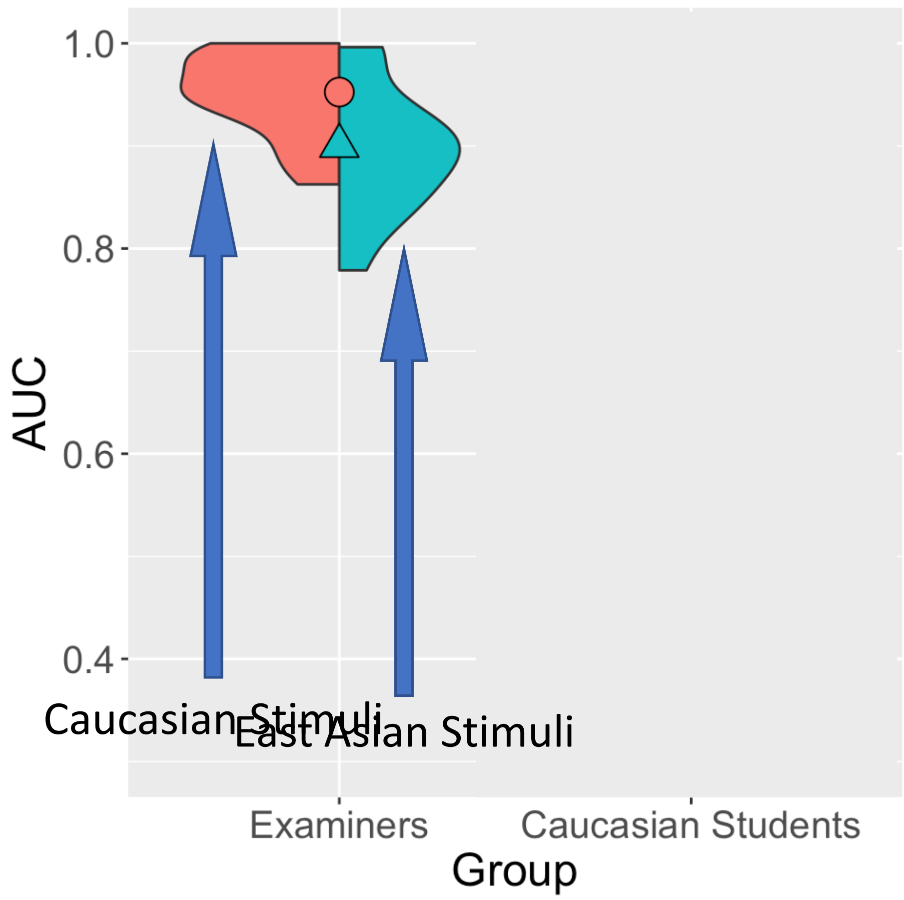
Cross-race test

NIST

- 80 pairs of face images
  - Caucasian faces
  - East Asian faces
- Up to 30 seconds to view each pair

- 14 Examiners (Caucasian ancestry)
- 48 Caucasian Students

Both stimuli sets



Caucasian Stimuli

- **Examiners > Caucasian Students (*p*=0.00123**)**

East Asian Stimuli

- **Examiners = Caucasian Students (*p*=0.123)**

Conclusions

- Perceptual tests
  - Limited time
  - No tools or methods
  - Black box needed
- Cross-Race Comparisons
  - Caucasian stimuli: examiners better than Caucasian students
  - East Asian stimuli: examiners equal to Caucasian students

Questions?

Thank you!